MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2018/2019

TMA 7021 – DATA MINING AND ANALYTICS

(TC01)

23 JANUARY 2019
10.00 a.m – 12.00 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of **6** pages with FOUR questions only.
2. Answer **ALL** questions.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

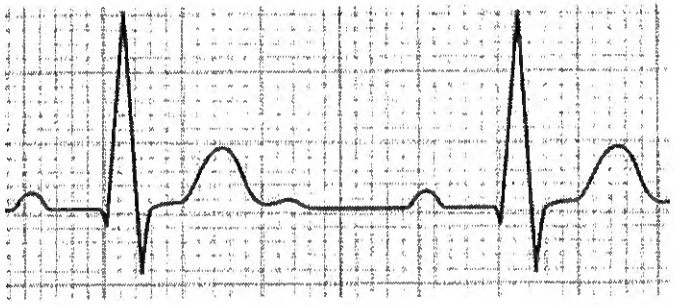
(a) Discuss TWO (2) characteristics of a domain where Bayesian Networks can be employed.

[2 marks]

(b) You are requested to design and draw a Bayesian Network to predict heart attack. You have been informed that the causes of heart attack are (i) smoking, (ii) stress, (iii) obesity, and (iv) lack of physical activity. You also know that symptoms of stress are (i) cold sweat, (ii) fatigue, and (iii) shortness in breath. The medical expert indicated that ECG can tell some information about stress.

[4 marks]

(c) Below shows an example of ECG pattern. Assuming that you have decided to discretize the ECG value, suggest ONE way to perform discretization.

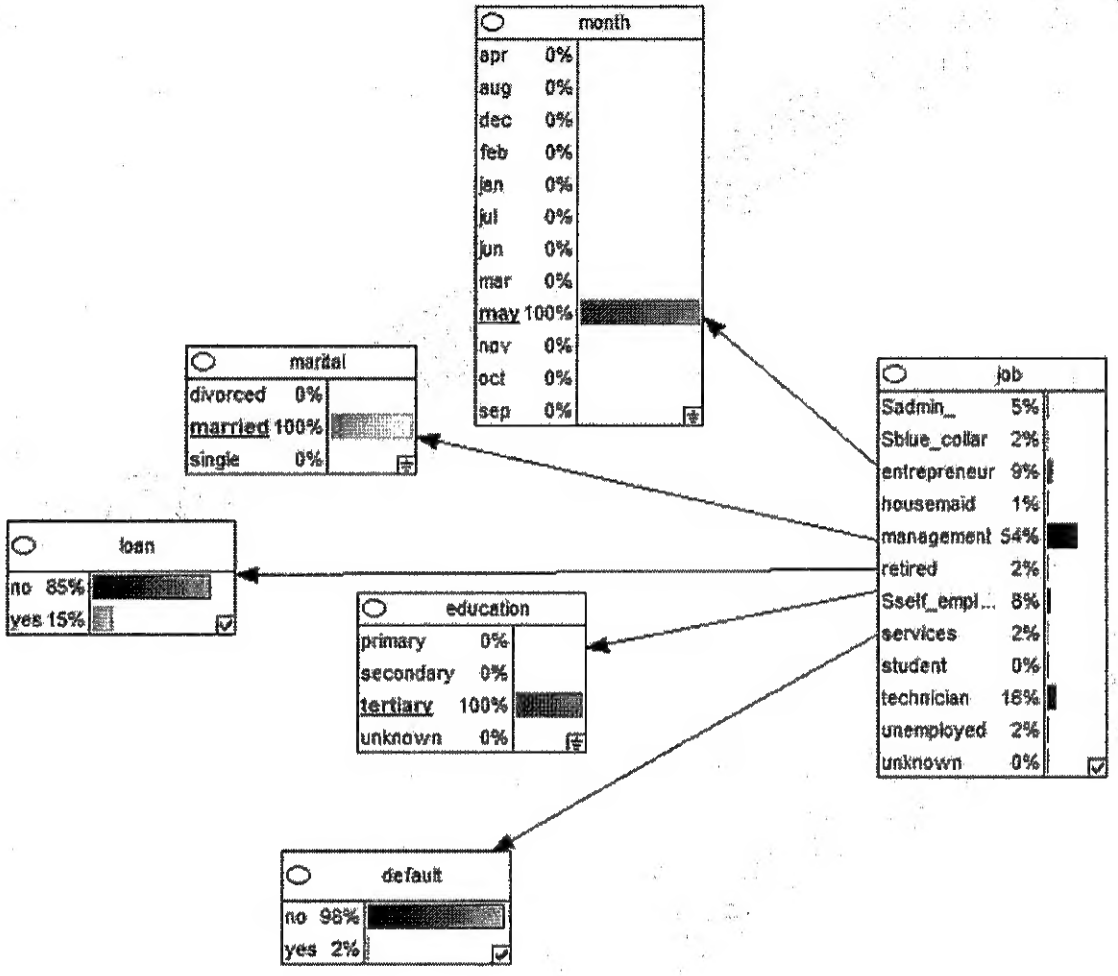


[2 marks]

Continued...

(d) Write the corresponding rule after evidence instantiation on the Bayesian Network below:

[2 marks]



Continued...

QUESTION 2

(a) Association rule mining and decision tree both generate rules. What is the main difference between the rules generated by the two techniques?

[2 marks]

(b) Insurance companies runs *cross-selling* campaign to get customers buy more products. That is, based on historical data, customers who purchased product A and B will likely to purchase product C. Suggest TWO (2) machine learning approaches to support cross-selling. Discuss the condition under which a particular machine learning approach can be applied.

[2 marks]

(c) The table below shows the fruits purchased by customers.

customer	Fruit
1	Apple, pineapple, durian
2	Durian, kiwi
3	Strawberry, kiwi
4	Apple, pineapple, durian
5	Pineapple, apple

(i) Calculate the *support* for the following item-set:

{strawberry, kiwi}

[2 marks]

(ii) Calculate the *confidence* for the following rule:

{durian, pineapple} → {apple}

[2 marks]

(iii) Calculate the *Lift* for the following association rules:

{apple, pineapple} → {durian}

[2 marks]

Continued...

QUESTION 3

(a) *Euclidean* and *Gower* are two different distance measurement methods. Study the dataset below carefully, name and discuss the method that can be used to calculate the similarity between any given two records. No transformation of the dataset is made.

location	**restaurant**	**school**	**population**	**average_salary**
ampang	yes	chinese	28000	5000
sedang	no	chinese	19000	4000
dengkil	yes	national	12000	3000

[2 marks]

(b) Study the following table carefully and answer the questions below.

Name	Average_age	Average_salary	population
ampang	25	3.2	89
dengkil	18	3.1	98
serdang	22	2.6	88

Calculate the similarity for (ampang and serdang) and (ampang and dengkil) using *Euclidean* distance measurement method. What can you conclude?

[5 marks]

(c) Discuss ONE difference between *k-means* and *hierarchical* clustering.

[3 marks]

Continued...

QUESTION 4

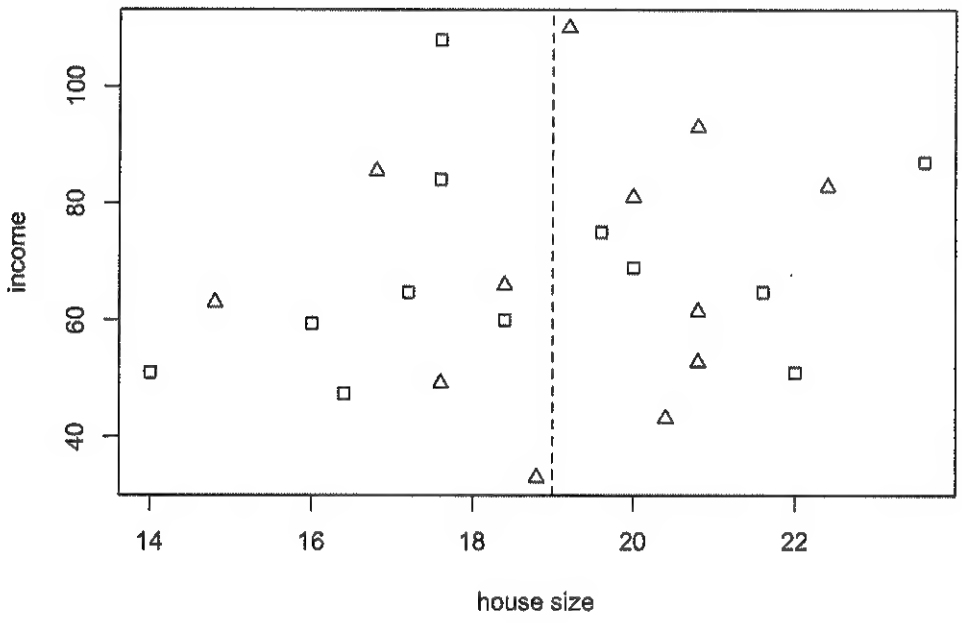
(a) k-Nearest Neighbour is a *lazy learner* while C5.0 is an *eager learner* while. Discuss ONE difference between the two learners.

[2 marks]

(b) Name TWO (2) methods to measure node impurity in a decision tree.

[2 marks]

(c) Scatterplot below shows the relationship between house size and household income. The triangles represent *owner* while the square boxes represent *non-owner*.



(i) Calculate the Gini Index of the entire dataset before splitting the dataset at house size = 19.

[2 marks]

(ii) Calculate the reduction in Gini Index by introducing a split of dataset at house size = 19.

[2 marks]

(iii) Write one possible rule generated based on the diagram above.

[2 marks]

End of Pages.